DALMOOC Week 8 - LightSide - Some Text Mining

14.12.14 / Ingrid Dethloff http://blog.idethloff.de

• LightSide's Example Dataset "sentiment_sentences"

• 10662 instances, two columns: class (neg/pos) and text

0		<u>.</u> 9 · (* ·)	÷	Mappe1 -	Microso	oft Excel	nichtkommerz	ielle Ver	wendung		-		x
	2	Start Einfüg	en Seitenlayout	Formeln	Daten	Überprü	fen Ansicht	Add-Ins	s		۷) _ (= x
Exte	erne Da brufen	Daten Alle Image: Construction of the constr								:ben 🔻	e line Ann		
		A1	▼ (* f _x										×
	А			В					С	D	E		F
1	class	text											
2	neg	simplistic, si	lly and tedious .										
3	neg	it's so laddish	and juvenile , only	teenage boys	could p	ossibly fir	nd it funny .						
4	neg	exploitative a	and largely devoid	of the depth or	sophist	ication th	at would make	watching	g such a graphic t	reatment o	f the crimes bea	rable	
5	neg	[garbus] disca	ards the potential f	or pathological	study,	exhumin	g instead , the s	kewed m	elodrama of the	circumstan	tial situation .		
6	neg	a visually flas	hy but narratively o	paque and em	notional	ly vapid e	xercise in style	and myst	tification .				
7	neg	the story is a	so as unoriginal as	they come , alr	eady ha	ving beer	n recycled more	times th	ian i'd care to coι	unt.			
8	neg	about the on	ly thing to give the	movie points f	or is bra	vado to	take an entire	y stale co	oncept and push	it through t	he audience's n	neat gi	rind
9	neg	not so much f	farcical as sour .										_
10	neg	unfortunatel	y the story and the	actors are serv	ed with	a hack scr	ipt.						_
11	neg	all the more	disquieting for its re	elatively gore-	free allu	isions to t	he serial murde	ers , but it	t falls down in its	attempts t	o humanize its	subjec	xt.
12	neg	a sentimenta	I mess that never ri	ngs true .									_
13	neg	while the per	formances are ofte	n engaging , th	nis loose	collectio	n of largely imp	rovised r	numbers would p	probably ha	ve worked bett	er as a	i on
14	neg	interesting,	out not compelling										_
15	neg	on a cutting r	oom floor somewh	ere lies foo	tage tha	at might h	ave made no si	ich thing	a trenchant , iroi	nic cultural	satire instead o	f a fru	stra
16	neg	while the en	emble player who	gained notice	in guy ri	tchie's lo	ck , stock and tv	o smokir	ng barrels and sn	atch has th	e bod , he's unli	kely to	o be
17	neg	there is a diff	erence between m	ovies with the	courage	e to go ov	er the top and r	novies th	at don't care abo	out being st	upid		_
18	neg	nothing here	seems as funny as	it did in analyz	e this , n	not even j	oe viterelli as d	e niro's r	ight-hand goom	bah.			_
19	neg	such master s	creenwriting come	s courtesy of jo	ohn pog	ue , the y	ale grad who pr	eviously	gave us the skul	Is and last	year's rollerbal	. enc	ough
20	neg	here, commo	on sense flies out th	e window , alo	ong with	the hail	of bullets , non	e of whic	h ever seem to h	it sascha .			_
21	neg	this 100-minu	ite movie only has	about 25 minut	tes of de	ecent mat	erial.						
22	neg	the execution	h is so pedestrian th	hat the most po	ositive c	omment	we can make is	that rob s	schneider actual	y turns in a	pretty convinci	ng per	TOR
23	neg	on its own , it	s not very interest	ing . as a remai	ke, it's a	a pare imi	tation.	ن ام م م ا	uill forming arrest	الم ما ما بر بم بر – ۲		anals :	. 1:44
24	neg	Taballa1	Some studios firmi	y believe that	peopler	lave lost	the ability to th	ink and w	viii iorgive any si	loady prod	uct as long as th	ere's a	
Ber	eit	Tabellet								III 100 9	× 🔾 🔲		(
											· · · ·		J

LightSide: Feature Extractor Plugins = Basic Features

01) Unigrams & Logistic Regression

Configure Basic Features = Unigrams & Include Punctuation & Track Feature Hit Location Build Models = Logistic Regression & Cross Validation 10fold

→ Result: 4485 Features // Accuracy = 0,759 // Kappa = 0,518

place Regults Compare Mod	LightSide			×
Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machi Decision Trees Weka (AII)	Configure Log	e Logistic Regression gularization gularization (Dual)		
Evaluation Options: Cross-Validation Supplied Test Set No Evaluation	Fold Assignment: Random By Annotation: By File Number of Folds: Auto Manual: 10	~		
Feature Selec	ction	Model C	onfusion Matrix:	8
Metric Accuracy Kappa	Value 0,759 0,5179	Act \P neg pos	red neg 4079 1318	pos 1252 4013
	plore Results Compare Mod Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Mach Decision Trees Weka (All) Evaluation Options: Cross-Validation Supplied Test Set No Evaluation Feature Select Model Evaluation Metrics: Metric Accuracy Kappa	LightSide plore Results Compare Models Predict Labels Learning Plugin: Naive Bayes Logistic Regression Linear Regression Linear Regression Decision Trees Weka (All) Evaluation Options: Fold Assignment: © Cross-Validation Image: Regression Supplied Test Set By Annotation: No Evaluation Image: Regression Quarter Selection Image: Regression Manual: 10 Image: Regression Quarter Selection Image: Regression Model Evaluation No Evaluation Image: Regression Image: Regression Test Set By File Number of Folds: Image: Regression Image: Regression Test Set Image: Regression	LightSide plore Results Compare Models Predict Labels Learning Plugin: Naive Bayes Image: Compare Models Image: Compare Models Incar Regression Support Vector Machines Image: Compare Models Image: Compare Models Decision Trees Decision Trees Image: Compare Models Image: Compare Models Image: Compare Models Evaluation Options: Fold Assignment: Image: Compare Models Image: Compare Models Image: Compare Models Image: Compare Model Test Set By File Number of Folds: Image: Compare Models Image: Compare Models Image: Compare Model Evaluation Image: Manual: 10 Image: Compare Model Compare Mode	LightSide plore Results Compare Models Predict Labels Learning Plugin:

02) Unigram, Bigram, Trigram & Logistic Regression

Configure Basic Features = Unigrams, Bigrams, Trigrams & Include Punctuation & Track Feature Hit Location

Build Models = Logistic Regression & Cross Validation 10fold

→ 02a) Whole Feature space of 12620 Features // Accuracy = 0,765 // Kappa = 0,530

Extract Features Restructure Data	Build Models Explore Results Com	oare Models Predict Labels			
Feature Tables:	Learning Plugin: Naive Bayes Linear Regr Support Vec Decision Tre Weka (All)	ession ssion tor Machines es		Configure Lo © L2 Regulariza O L1 Regulariza O L2 Regulariza	gistic Regres ation ation (Dual)
Class: class	Evaluation Optic	ns: Fold Assignment:	0 Max		
Train Name: 0	git_123grams_1 Feat	ure Selection			
Trained Models:	Model Evaluation	n Metrics:	Model Con	fusion Matrix:	
logit_123grams 🗸 📔	Metric	Value	Act \Prec	d neg	pos
TRAINED_MODEL Documents: sentiment_sente TRAINED_MODEL Documents: sentiment_sente TRAINED_MODEL TRAINED_MODEL Document_senters TRAINED_MODEL TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED TRAINED T	nces.csv	0,7651 0,5301	pos	4083 1257	1248 4074

0		LightSide				- 🗆 🗙	
Extract Features Restructure Data Build Models Ex	plore Results Compare Models	Predict Labels					
Feature Tables:	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machine Decision Trees Weka (All)	s		Configure Logistic Regression L1 Regularization L2 Regularization L2 Regularization (Dual)			
-Type: nominal	Evaluation Options: Cross-Validation Supplied Test Set No Evaluation	Fold Assignment:	10				
Name: logit_123grams_	2 Feature Selection	in #: 3500				8	
Trained Models:	Model Evaluation Metrics:			Model Co	onfusion Matrix:		
Incluster in the interview of the interv	Metric Accuracy	Value 0,7683		Act \Pr	red neg 4072	pos 1259 4120	
Documents: sentiment_sentences.csv Feature Plugins: basic Feature Table: 123grams Learning Plugin: Logistic Regression Wrapper Plugins: select Validation: CV V		0,000			46.44	120	

→ 02b) Feature Selection -> 3500 Features // Accuracy = 0,768 // Kappa = 0,537

03) Model Comparison Unigrams / Unigrams, Bigrams, Trigrams

Number 01: Unigrams & Logistic Regression (All Features) Number 02b: Unigrams, Bigrams, Trigrams & Logistic Regression & Feature Selection = 3500

→ Significant Improvement (p=0,014*, t=2,468)

Model 01 predicted 1252 as positive and 1318 as negative when the data said the opposite Model 02b predicted 1259 as positive and 1211 as negative when the data said the opposite

۹			Ligh	tSide			- 🗆 🗙		
Extract Features Restr	ucture Data Build Models	Explore Results Compare Mod	els Predict	Labels					
Baseline Model:				Competing Model:					
logit1grams		✓	×	logit_123grams_1 V					
TRAINED_MODEL TRAINE	ntiment_sentences.csv s: basic 1grams : Logistic Regression logit1grams 9		TRAINED_MODEL Documents: sentiment_sentences.csv Feature Plugins: basic Feature Table: 123grams Learning Plugin: Logistic Regression Wrapper Plugins: select Validation: CV FoldMethod: AUTO -numFolds: 10 Validation: CV 						
Comparison Plugin: B	asic Model Comparison						~		
Baseline Model Metric Metric Accuracy Kappa	Baseline Model Metrics: Metric Value Accuracy 0,759 Kappa 0,5179				Competing Model Metrics: Metric Value Accuracy 0,7683 Kappa 0,5367				
Baseline Confusion Ma	atrix:			Competing Confusion Matrix:					
Act \Pred	neg	pos		Act \Pred	neg	pos			
neg	4079	1252		neg	4072	1259			
				Significant improvement	t (p=0,014*, t=-2,468)				
Get Support					r Multit	threaded 0.3 GB us	ed. 1.0 GB max 🗎		
C ber support									

LightSide: Feature Extractor Plugins = Basic Features & Stretchy Patterns

04) Unigrams & Logistic Regression

01a) Configure Basic Features = Unigrams & Include Punctuation & Track Feature Hit Location Configure Stretchy Patterns = (default Pattern Length=2-4 / Gap Length = 1-2) // Add LightSide Categories negative.txt and positive.txt) // check off "Require at least one category per pattern"

-> Result: 5134 Features



04b) Build Models = Logistic Regression & Cross Validation 10fold

→ Result: 5134 Features // Accuracy = 0,759 // Kappa = 0,517

04c) Build Models = Logistic Regression & Cross Validation 10fold & Feature Selection = 3500 → Result: 3500 Features // Accuracy = 0,767 // Kappa = 0,535

(3)		LightSide			_ 🗆 🗙
Extract Features Restructure Data Build Models Exc	olore Results Compare Mode	els Predict Labels			
Extract Features Restructure Data Duilo Model Exp Feature Tables: Stretch_1grams FEATURE_TABLE Documents: sentiment_sentences.csv Feature Plugins: stretch_basic Feature Table: stretch_1grams -5134 features -Gas: class: Type: nominal	Iore Results Compare Mode Learning Plugin: Naive Bayes Logistic Regression Linear Regression Decision Trees Weka (All) Evaluation Options: Cross-Validation Supplied Test Set No Evaluation	els Predict Labels	Max	Configure © L2 Regu O L1 Regu O L2 Regu	Logistic Regressio
Name: logit_stretch_1gra	ms_2 Feature Select	tion #: 3500			۵
Trained Models:	Model Evaluation Metrics:		Model C	onfusion Matrix:	
logit_stretch_1gra 🗸 📙 🗙	Metric	Value	Act \P	red neg	pos
TRAINED_MODEL	Accuracy	0,7674	neg	4131	1200
Documents: sentiment_sentences.csv Feature Flugins: stretch basic Feature Table: stretch_Igrams Learning Plugins: cojistic Regression Wrapper Plugins: select Validation: CV foldMetbod: ALITO		JU,5348		1280	4051
🚱 Get Support			-¢1	Multithreaded	0,4 GB used, 1,0 GB max 🗒

05) Model Comparison Unigrams / Unigrams, Stretchy Patterns

Number 01: Unigrams & Logistic Regression (All Features) Number 04c: Unigrams & Stretchy Patterns & Logistic Regression & Feature Selection = 3500

→ Highly Significant Improvement (p=0,002*, t=3,14)

Model 01 predicted 1252 as positive and 1318 as negative when the data said the opposite Model 04c predicted only 1200 as positive and 1280 as negative when the data said the opposite

0		Ligh	ntSide		- 🗆 🗙			
Extract Features Restructure	Data Build Models Explore	Results Compare Models Pred	ict Labels					
Baseline Model:			Competing Model:		_			
logit1grams		✓ <a> 	logit_stretch_1grams_1 V					
TRAINED_MODEL Documents: sentimen Feature Plugins: basis Learning Plugin: Logis Validation: CV Validation: CV Accuracy: 0,759 Kaopa: 0,518	t_sentences.csv : is stic Regression _1grams	~	TRAINED_MODEL Documents: sentiment_sentences.csv For Feature Plugins: stretch basic For Feature Table: stretch_Igrams For Learning Plugin: Logistic Regression For Wrapper Plugins: select FoldMethod: AUTO FoldMethod: AUTO FoldMethod: 10					
Comparison Plugin: Basic M	odel Comparison				¥			
Baseline Model Metrics:			Competing Model Me	trics:				
Metric	Value		Metric	Value				
Accuracy	0,759		Accuracy	0,7674				
Baseline Confusion Matrix:			Competing Confusion	n Matrix:				
Act \ Pred	peq	pos	Act \ Pred	290				
neg	4079	1252	Deg.	4131	1200			
pos	1318	4013	DOS	1280	4051			
			Highly significant impro	vement (p=0,002**, t=-3,14)				
6 Get Support				载 Multithreaded	0,3 GB used, 1,0 GB max 📋			

06) Explore Feature Space Model 04c: Unigrams & Stretchy Features & Feature selection=3500

Configure Confusion matrix select: Data negative & Prediction Positive = 1200 Configure "Evaluations to Display": check off "Frequency" and "Feature Weight" // Sorting by Frequency

Configure "Exploration Plugin" = Documents Display // check off "Filter Documents by selected feature" and "Documents from selected cell only"

0		Ligh	ntSide					×
Extract Features Restructure Data Build Models Explo	ore Results Comp	are Models Pred	ict Labels					
Highlight:	Cell Highlight:	ell Highlight: 😽 Fe			Features in Table:			•
logit_stretch_1grams_1 🗸 📙 🗙	Act \Pred	Act \Pred neg pos			Search:			
TRAINED_MODEL	neg 4131 0 1200			Feature	Frequency	Feature Weight		
Documents: sentiment_sentences.csv	pos	0 1280	0 4051		<pre>O <period></period></pre>	1144	?	~
Feature Plugins: stretch basic					<comma></comma>	742	0,1756	
Feature Table: stretch_1grams					O the	724	?	
Learning Plugin: Logistic Regression					O a	613	0,1258	
Wrapper Plugins: select					() and	545	0,3116	
CV Subation: CV					Oof	530	0,0778	
minimetriod: AUTO					O to	390	-0,1461	
numFolds: 10					() 's	376	0,0485	
Trained Model: logit_stratch_logame_1	Evaluations to	Display:				352	?	-
	Horizontal	Difference		^		346	0,1006	-
	Vertical Ab	solute Difference				2/8	0,1534	-
	Vertical Dif	ference			Obut	277	r -0.1466	-
	Model Analysis				film	201	0,1100	-
	Feature In	fluence			with	181	0,2000	-
	Feature Se	lection			for	179	-0.0892	-
	Feature W	eight		¥		166	0,0092	¥
✓ Filter documents by selected feature Reverse document filter ✓ Documents from selected cell only Instance Predicted Actual Text ✓ 1 pos neg it's so ladd 12 pos pos neg 32 pos neg payami tri 39 pos 12 pos neg curling ma 32 pos neg curling fma 39 pos pos neg the effort 55 pos neg an odd , h 55	Insta it's nd_i	nce 1 (Predi so laddish t funny .	cted pos, Actual	l neg) nly teenage	boys could	possibly fi 🔷	
56 pos neg though he 57 pos neg pascale b								
79 pos neg takes one	v							
Get Support					🔩 Mu	ultithreaded	0,3 GB used, 1,0 GB ma:	x 📋

Task: Use this interface to explore which features got the most weight in your model. It's most important to consider features that both got a lot of weight and occurred more than just a couple of times. Which features were most important? What did the stretchy pattern features add?

Data negative & predicted positive

 As I included punctuation in the Basic Features Extraction, these (period, comma) have the highest frequency and lowest or no weight at all. Other features with a high number for false positives (frequency /weight) are: "you" (=135 / 0,54 -> normally positive term), "n't" (103 / - 0,476 -> normally negative term), "all" (67 / -0,333 -> normally negative term), "what" (58 / 0,375 -> normally positive term), "so" (50 / 0,361 -> normally positive term), "way" (40 / 0,148 -> normally positive term), "love" (37 / 0,441 -> normally positive term) The feature "n't" as a negative form would be associated with negative sentiment and when you check this feature in the confusion matrix for "data negative & prediction negative", you get a high frequency of 456.

• Stretchy patterns added context, for example the feature "STRONG-POS [GAP] but" (9 times), which puts into perspective a positive term: In the original text, this would fit to "...good intentions, but", "...great team, but".

The other way round, if you look at "data negative & predicted negative" in the confusion matrix, there are more features of this kind: "STRONG-POS [GAP] but", "STRONG-POS [GAP] but the", "STRONG-POS [GAP], but", "STRONG-POS [GAP]. but" etc. The feature "but" which stands alone then has a weight of 0,147 which is so small that it indicates (in my understanding), that from this word alone, you can't predict if something is meant positive or negative.